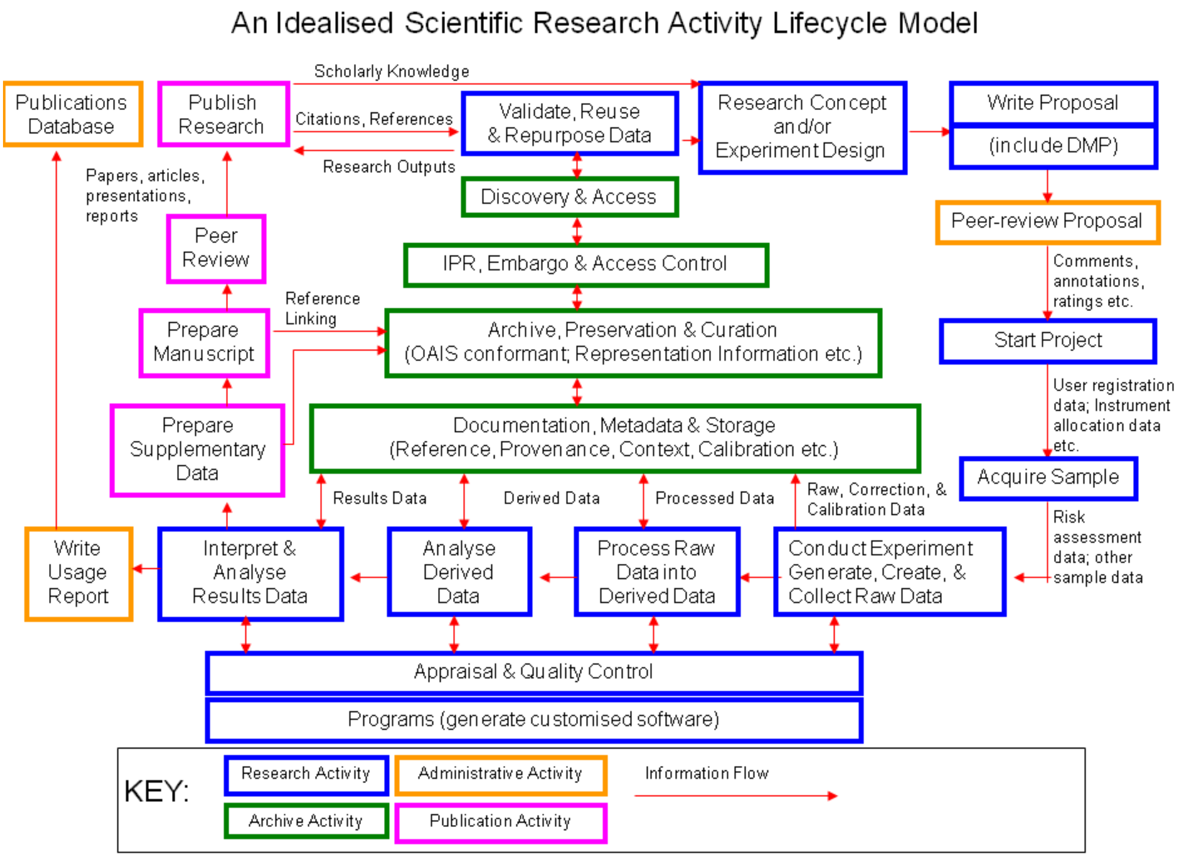
I2S2 Idealised Scientific Research Activity Lifecycle Model

The model represents the processes and phases of a typical physical science experiment project as indicated in the blue text boxes. The stages include: development of the research proposal; its peer-review; carrying out of the experiment; and processing, analysis and interpretation of the data which is eventually reported and published in various forms as research outputs. In addition to these familiar phases, we have incorporated several idealised stages (green boxes) to cater for the long-term management and availability of scientific data; these include appraisal and quality control; documentation including metadata and contextual information; storage, archive, preservation and curation; and IPR, embargo and access control.



It is important to realise that the life expectancy of scientific data has increased over the years as more and more scientific research becomes derivative in nature, dependent on data generated, managed and made widely accessible to third parties. However, effective reuse and repurposing of data requires much more information than the dataset alone. Trust and a thorough understanding of the data is a precursor to its reuse and this in turn necessitates transparency and access to considerable contextual information regarding how the data was generated, processed, analysed and managed.

Consequently, within the I2S2 project, research data is considered to be not only the raw images and numerical datasets that are generated and collected from scientific experiments, but also the broader categories of information that are associated with such data. These include research and experiment proposals; the results of the peer-review process; laboratory notebooks; equipment configuration and calibration data; processing software and associated control parameters; wikis; blogs; metadata (context, provenance etc.) and other documentation necessary for the interpretation and understanding of the scientific data (semantics); as well as the administrative and safety data that accompany experiments (from the yellow Administrative Activity boxes).

Much of this type of information can be considered to be Representation Information (RI), a key concept which underlies the Open Archival Information System (OAIS) Reference Model [1]. RI is a very broad concept, encompassing any information required to render, process, interpret, use and understand data. For example, it may include a technical specification, or a data dictionary exposing semantics or even a software processing tool. An investigation of RI for chemical crystallography data was undertaken in the eCrystals Federation Project [2].

The data and information flows between and into the various phases of the research cycle are represented by the arrows between the boxes. Information generated during the development of the research concept and experiment design is most likely to be textual and paper-based in the form of hand-written notes or hand-drawn diagrams. An electronic record is likely to emerge at the proposal writing stage in the form of a document, or a collaborative tool such as a wiki or blog (or even email). The peer-review process generates additional (most likely textual) information, but also ratings of research proposals and possibly funding and resource allocation data.

Before an experiment can take place, administrative and sample safety information is required to be recorded in the form of either digital or analogue forms which must be checked and approved. An important part of the experiment setup phase is the recording of equipment configuration and calibration information, which is often generated automatically by the instrument being used. Whilst the experiment is in progress, the scientist is likely to record additional information in an analogue laboratory notebook or a laptop computer.

Following the collection of data from the experiment, there is normally a stage in which the raw datasets are cleaned and checked to make them usable, resulting in processed data. Once again, it is likely that the scientist will record any observations or issues in a notebook or laptop.

Results data are produced from an iterative cycle of processing and analysing derived data using software applications and tools. This is the stage in which the real scientific work is performed and it is crucial that adequate records are kept for future reference; most researchers record such information in a laboratory notebook or a laptop.

The traditional end to a scientific research experiment or study is the writing up and publishing of the results in a conference or journal article, with the results data distilled and interpreted to the extent that they cannot be easily checked or verified.

Until recently, it has been the norm that scientists are prepared to share the results dataset with selective colleagues. However, with improvements in technology there is an increasing demand to make available raw, processed and derived data for validation and reanalysis purposes, necessitating data management of these types of data as well as the results data.

Documentation of all types of scientific data (raw, reduced, processed, derived and results) in terms of providing adequate metadata, contextual information and Representation Information, becomes very important for their maintenance and management as well as for additional purposes such as: referencing and citation; authenticity; integrity; discovery and access; search and retrieval; reuse and repurposing; preservation and curation; IPR, embargo and access management. All of these functions support the scholarly research and communications process.

References

1. Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System*, ISO:14721:2002, (2002), http://public.ccsds.org/publications/archive/650x0b1.pdf#search=%22OAIS%20model%22
2. Manjula Patel, Simon Coles, David Giaretta, Stephen Rankin, Brian McIlwrath, *The Role of OAIS Representation Information in the Digital Curation of Crystallography Data*, IEEE eScience 2009, Oxford, UK, 9-11th December 2009